ELSEVIER

Applying computational modeling to drug discovery and development

Neil Kumar, Bart S. Hendriks, Kevin A. Janes, David de Graaf and Douglas A. Lauffenburger

Department of Chemical Engineering, Pfizer Research Technology Center and Department of Biological Engineering MIT, Cambridge, MA 02139, USA

Computational models of cells, tissues and organisms are necessary for increased understanding of biological systems. In particular, modeling approaches will be crucial for moving biology from a descriptive to a predictive science. Pharmaceutical companies identify molecular interventions that they predict will lead to therapies at the organism level, suggesting that computational biology can play a key role in the pharmaceutical industry. We discuss pharmaceutically-relevant computational modeling approaches currently used as predictive tools. Specific examples demonstrate how companies can employ these computational models to improve the efficiency of transforming targets into therapies.

Introduction and motivation for use of models

Prediction is the attempt to use existing knowledge to foretell an event before it happens. Whether it is a biologist trying to predict how target inhibition will affect cell behavior, a physician trying to predict how a drug will affect a patient, or a manager trying to predict future return on investment, prediction plays a vital role in the pharmaceutical industry. Breakthroughs in the fields of genetics, biochemistry and molecular biology have increased our ability to understand and predict behavior in biological systems. Here we argue that computational modeling based on biological information can be used to extend the limits of our understanding, thereby enhancing predictive accuracy. Models used to simulate cellular or human biology produce reliable data and new hypotheses, and can translate information between *in vitro* screens, cell-based assays and ultimately patients. This extension of knowledge is valuable to the pharmaceutical industry for novel product generation. Even in the absence of novel products, incorporation of computational modeling in the ways outlined below could save millions of dollars based on increased efficiency [1,2].

A coarse-grained schematic of the pharmaceutical research and development (R&D) pipeline is shown in Figure 1. We have identified three areas where computational modeling has potential to substantially impact efficiency and development. The first area is cell-signal behavior, where the application of models characterizes how lead compounds affect intracellular signaling. The second area is signal-response behavior, where models predict cellular phenotype from signaling information. The third area is physiology, in which models are used to simulate clinical outcomes. Each class of model can help identify new drug targets. We address each application area separately, highlighting important work relevant to the pharmaceutical industry.

In addition to specific applications, there is also a natural role for modeling to link traditional biology and high-throughput informatics analysis (Figure 2). For example, the construction of a signaling model begins with an assembly of molecular interactions, rate parameters and spatial restrictions. Informatics groups analyze high-throughput datasets (ie. gene–chip arrays, gene sequencing results, mass spectrometry results and yeast two-hybrid results) using methods like clustering or spacing alignments, and integrate results with data from other in-house biological experiments and from literature (obtained by text mining). The data are then further organized into ontologies [3]. A model is constructed from a subset of these data and is then validated using traditional biology experiments. If the model captures experimental trends, it is used to generate predictions or hypotheses that suggest new biological experiments. The results of these experiments either further validate the model or identify novel biology that is then incorporated into the model. This interplay between

Corresponding author: Kumar, N. (nkumar@mit.edu)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.07.010

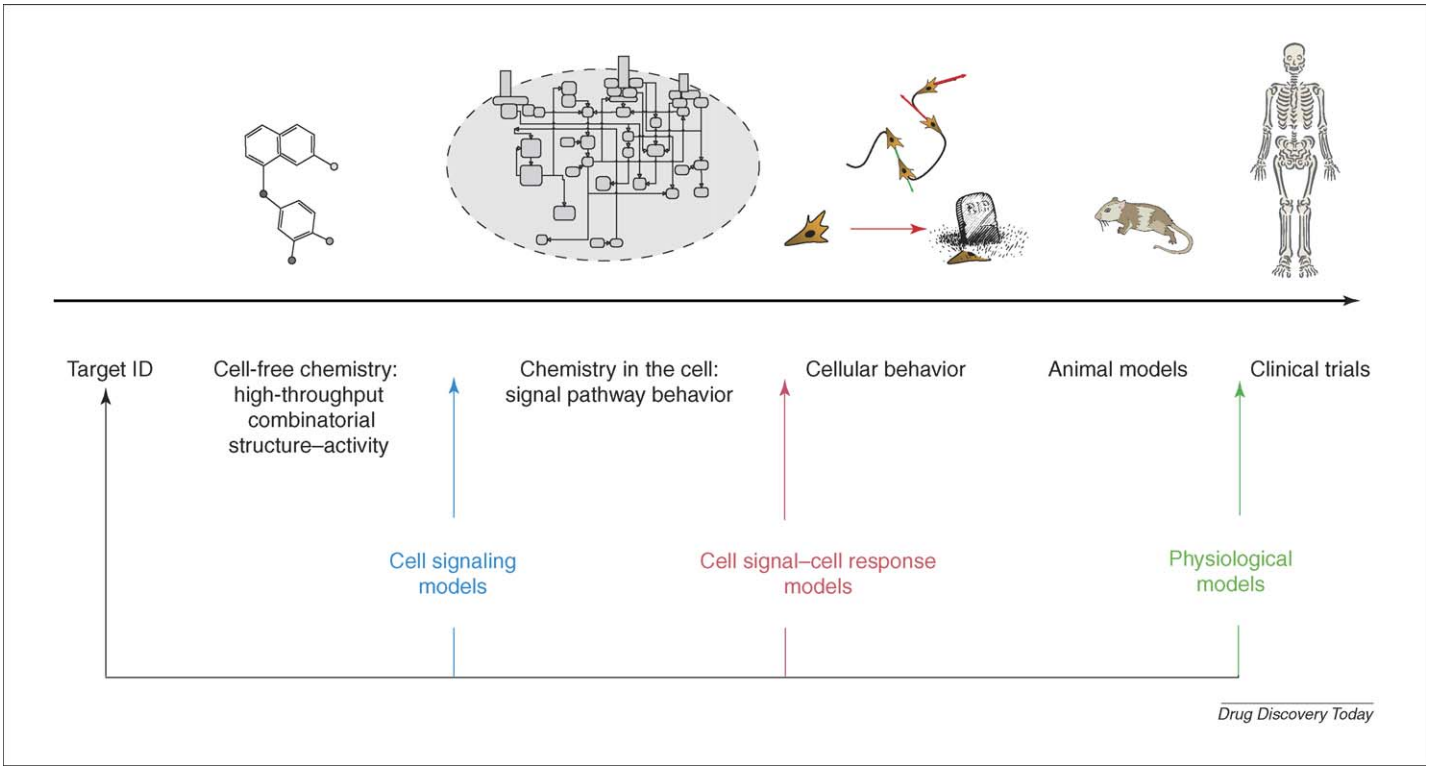

FIGURE 1

Areas of impact for computational modeling in the pharmaceutical R&D process. A course-grained diagram of the R&D process illustrates three potential areas for model application. Cell signaling models simulate intracellular signaling dynamics and predict drug effects on signaling. Cell signal–cell response models correlate intracellular signals to cell behaviors, such as migration and apoptosis. These models predict drug effects on cell behavior. Physiological models simulate organ level behavior and predict human response to drugs. All three model classes could additionally reveal novel drug targets.

informatics, modeling and traditional biology enables the focused use of large datasets to solve biologically relevant problems.

We have restricted the scope of our discussion to selected modeling efforts with emerging relevance to the pharmaceutical industry. We do not, however, address many models already used successfully by the pharmaceutical industry. For example, pharmacokinetic (PK) models, which we do not cover here, are perhaps the most significant class of models being used today. In addition, we do not discuss the myriad other metabolic models that have been successfully employed at cellular and animal levels. Nevertheless, we hope to demonstrate how new advances in computational modeling can integrate into the R&D workflow, and in so doing we attempt to establish a framework for thinking about the application of diverse types of models to the pharmaceutical industry.

Cell signaling models

Defects in signal transduction underlie many diseases that are of interest to pharmaceutical companies. For example, dysregulation of conserved protein tyrosine kinase pathways leads to a variety of cancers [4]. Individual signaling proteins inside the cell are often the target of small-molecule drugs, whereas many antibody drugs target the receptors controlling signaling cascades. Here, we highlight a group of computational models that have been used to describe signaling pathways relevant to disease. These models, in comparison to the signal-response models discussed in the next section, are more highly specified in their molecular details [5]. Typically, ordinary differential equations (ODEs) are used to describe mass-action kinetics and system behavior. Experimental measurement of reaction rates, concentrations, molecular interactions and trafficking parameters are essential for the construction of such models. The level of detail necessary varies from system to system, but many signal-transduction pathways can be modeled using a combination of measured values, fitted parameters, and coarse-grained descriptions of interactions [6].

Models that describe signaling pathways are important in pharmaceutical research for three main reasons: (i) they often capture nonintuitive signal behavior and identify novel molecular function; (ii) they allow researchers to experiment *in silico* across a wide range of conditions (e.g. receptor numbers, ligand concentrations and phosphorylation rates), thus saving experimental resources and identifying important further experiments; and (iii) they serve as a database for much of the known information about a particular pathway.

Lee *et al.* [7] used ODE-based models together with experimental data to analyze the Wnt signaling pathway, a crucial pathway in the physiology of diseases such as cancer. Model construction from published data revealed that further experiments to measure total concentrations and rates of dissociation were necessary. Successful implementation of the model showed that adenomatous polyposis coli (APC) and axin, both of which coordinate the degradation of the Wnt pathway effector β-catenin, bind proteins in different ways, with axin binding randomly and APC binding in an ordered manner to promote degradation. The model also highlighted APC-dependent axin degradation, a mechanism potentially important in explaining β-catenin levels in response to APC

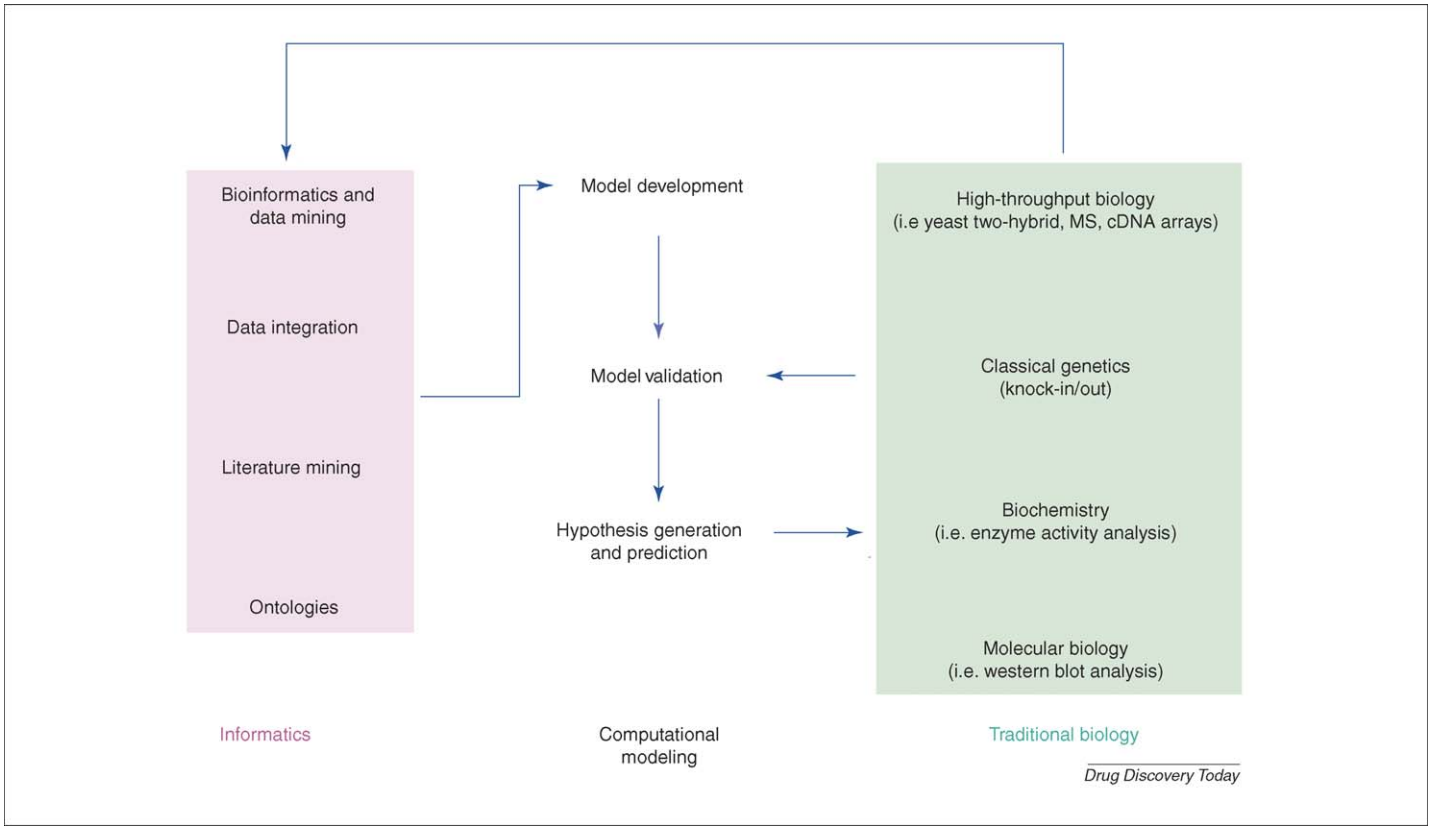

FIGURE 2

Computational modeling in the R&D workflow. Computer models rely on data from informatics and traditional biology applications. As such, they help coordinate the use of informatics to answer biological questions. Informatics groups analyze large biological datasets, mine scientific literature for data, integrate disparate forms of data and organize databases for further use. Computational models are developed using subsets of stored data. Models are validated against results from genetic, biochemical, or molecular biology experiments. After validation, models are used to predict and interpret novel biology. Resulting hypotheses are tested using traditional biology techniques. If predicted data do not correspond to experimental measurements, the model is altered to incorporate the new biological information.

mutation. Using the model, the effects of perturbing seven different Wnt pathway molecules were explained, exemplifying the potential target selection and drug testing applications of such an approach. In addition, the relevant investment of this modeling compared with exclusive experimental testing is considerably smaller. Model predictions also indicated that slight changes in APC function lead to significant buildup of β-catenin levels. In colorectal cancer, APC mutations give rise to elevated β-catenin levels, resulting in a cancerous phenotype. Potential therapies to this based on modeled proteins might include drugs that inhibit protein phosphatase 2A (PP2A) or break down T-cell factor. The model indicates that PP2A inhibition would be more effective in terms of β-catenin degradation, but this would also affect axin levels to a greater extent, which might in turn perturb other signaling pathways reliant on axin [7,8].

Hendriks *et al.* [9] developed a detailed model of ErbB receptor signaling dynamics coupled to a receptor trafficking model. ErbB receptors and their signaling pathways are implicated in various cancers and are the target of many therapeutic compounds [4]. The model was developed using binding, dimerization and trafficking rate constants from literature. Hendriks *et al.* simulated time-resolved phosphorylation profiles for three types of ErbB receptors under more than ten ligand stimulating conditions. These simulations required less than24 h on a desktop computer; by comparison, the equivalent experimental data would have required weeks of effort and resources at the bench. Further simulation produced phosphorylation profiles under varying assumptions, including dephosphorylation of receptors at the surface versus exclusive dephosphorylation of receptors internalized in the endosome. Comparison with experimental data revealed that ErbB receptor dephosphorylation occurred primarily in the endosomal compartment. Although many drugs target the extracellular domain of ErbB receptors to prevent ligand binding and subsequent phosphorylation, the model predicts that shunting more receptors per unit of time to the endosome will attenuate phophorylation. Inhibition of Sprouty, a class of molecules that are known to inhibit receptor internalization through interaction with Casitas B-lineage (Cbl), could increase trafficking to the endosome, thus providing a molecular target for testing this model-derived hypothesis [10]. In addition, scientists could use the model to predict the degree of elevated trafficking needed to achieve a given reduction in ErbB phosphorylation levels. But how much reduction in phosphorylation is enough to affect downstream signaling? Again, computational modeling can help. Well known downstream effectors of the ErbB pathways include extracellular-regulated kinase (ERK) and protein kinase B (PKB or AKT). Hatakeyama *et al.* [11] and Schoeberl *et al.* [12] have constructed models that simulate the effect of receptor phosphorylation on these important signaling intermediates. The models identify how variation in total amounts of phosphorylation and rates of

phosphorylation affect downstream signaling. Importantly, they also reveal how signaling molecules in distinct, but connected, pathways are regulated after ligand activation or signal inhibition.

All of the above efforts were effective because they were validated using experimental methods, they explored many more hypotheses than would be experimentally feasible and they highlighted nonintuitive but important regulatory schemes for signaling. Recent documented efforts indicate that cell signaling models are being successfully employed by pharmaceutical companies. For example, AstraZeneca recently used a computational approach to link the efficacy of the cancer drug Iressa to impaired receptor internalization and reliance on downstream AKT signaling [13]. The further application of these types of models relies on clear relationships between intracellular signaling proteins and cell behavior or higher-level processes. The general approach suffers when these relationships are poorly defined. The signal-response models we introduce next surmount this difficulty to predict complex cellular behavior.

Signal–response models

The cells that make up the human body engage in a large number of behaviors: cell migration, differentiation, proliferation and apoptosis are just a few of the many functions cells carry out. Interestingly, it has been hypothesized that no more than 20 signal transduction cascades control the seemingly endless list of cell behaviors observed in humans [7]. How does the cell, then, use these 20 or so cascades to coordinate cell behavior? Part of the answer is that each pathway can be activated in quantitatively different ways. For example, in PC12 pheochromocytoma, transient activation of the mitogen-activated protein kinase (MAPK) pathway leads to proliferation whereas sustained activation of this exact same pathway leads to differentiation [14]. Another part of the answer lies in the fact that multiple pathways can be used to control one behavior. For example, hepatocyte growth factor (HGF)-stimulated neuronal migration is regulated both by the MAPK and phosphatidylinositol 3-kinase (PI3K) signaling pathways [15]. Both of these facts suggest that to correct aberrant cellular behavior with drugs requires quantitative knowledge about multiple signaling proteins (that is, multivariate datasets). Multivariate datasets can then be used to understand cellular decision-making processes in the context of computational models.

The signaling pathway models discussed earlier give us insight into how the MAPK pathway might be manipulated to yield either transient or sustained ERK activation. However, mapping of this activation onto cellular behavior is an arguably more difficult task. Whereas ODE-level models are becoming more prevalent for describing signaling pathways, there are very few models that can accurately connect signaling pathways to cellular behavior at this level of mathematical description. The problem, therefore, requires the use of more abstracted signaling models [5]. Abstracted models identify statistical relationships between signals and behavior, which suggest causal signal–behavior relationships that can be further probed using molecular biology or genetic approaches. Here, we address a few recent models that have used multivariate cell-signaling data to reveal governing principles of cellular behavior. Predictions made on the basis of these models can reveal how a drug, or class of drugs, will affect a given cellular behavior.

Many recent efforts have focused on using modeling to identify genomic and proteomic groups of molecules responsible for disease-relevant cellular behaviors. For instance, Zheng *et al.* [16] used cDNA microarrays together with 2D gel electrophoresis and mass spectrometry to study the molecular effects of an acute promyelocytic leukemia (APL) cell line treated with retinoic acid (RA) and arsenic trioxide (ATO). The question being asked was: how do downstream signaling events coordinate a known program of differentiation and apoptosis? The answer to this question could shed some light on the signaling events responsible for clinical efficacy in APL patients co-treated with RA and ATO in conjunction with chemotherapy. Zheng *et al.* used a computational technique called self organizing maps (SOMs) to cluster signaling data and then characterized sets of signals important for the differentiation-versus-apoptotic cell program. They found that activation of the interferon and calcium signaling pathways coordinated differentiation and apoptosis. They also identified several transcription factors important for coordinated cell behavior. By isolating sets of genomic and proteomic changes associated with a specific therapeutic result, the authors identified drug mechanism and revealed important sets of pathways for future drug development. Conceptually similar studies have measured quantitative proteomic and genomic data, and then used computationally-aided organization of this data to interpret the role of signaling groups on cellular response [17,18]. Recent efforts establishing novel ways to integrate multiple types of measurement datasets should enable future successes for the application of this approach [19,20].

Although the previously mentioned signal–response studies rely on computation to identify co-regulation of signals, their goal is generally one of classification (also referred to as profiling or signature analysis) rather than prediction. To predict, quantitative measurement of the signal and/or transcription state and cellular response is required. We now discuss recent efforts that have established computational models to understand cellular behavior in the context of cell signaling. Models of this type are very useful for the pharmaceutical industry, as they allow scientists to alter signals and then predict how cellular behavior changes *in silico*. Janes *et al.* [21] describe a procedure based on linear modeling (partial least squares regression), whereby ~8000 intracellular signals were correlated with >1000 apoptosis-related cellular responses. The model computationally organizes the vast amount of signaling data measured, just as in previously mentioned studies, but takes the analysis a step further by deriving a set of parameters that map the signaling values onto the apoptosis measurements. Importantly, Janes *et al.* tested their model by perturbing the cellular signal state and then comparing experimentally-measured apoptotic values with those predicted from the model. For example, C225, an antibody raised against epidermal growth factor receptor (EGFR), was used to treat a human colon carcinoma cell line. Excellent agreement between predicted and measured levels of apoptosis validated the model and suggests that it can be used to understand apoptotic response under a wide range of signal perturbations (induced by drugs or cytokines). Indeed, the humanized form of C225, known as cetuximab, was approved by the FDA for treatment of advanced colorectal cancer, demonstrating that these approaches have direct pharmaceutical relevance. Furthermore, a recent published effort between Pfizer and

academic researchers [22] shows that signal-response modeling is already being applied successfully within the context of the pharmaceutical industry. In this study, Clayton *et al.* created a linear model, using NMR-measured pre-dose metabolite profiles, to predict the effect of paracetamol (acetaminophen) on rat excretion profiles and liver damage [22]. In addition, we recently applied this modeling technique to understand the effects of ErbB2 receptor overexpression (implicated in a large number of breast cancers) on cell proliferation and migration in human breast epithelial cells (unpublished), suggesting that this technique is broadly applicable to disease-relevant cell systems.

Sometimes, for systems that have been studied in great detail, it becomes possible to create a more mechanistic model of cell behavior based on intracellular signals. Gene Network Sciences has developed models for human cell proliferation and apoptosis to study potential anticancer strategies. Using a combination of known protein interactions and network inference, Christopher *et al.* [23] describe a model based largely on differential equations that predicts cell proliferation. While highly promising, it is worth noting that deterministic modeling in the area of cell behavior has proven very difficult, in part because of the level of detail required. This fact prompted GNS to develop abstracted inference models to connect their deterministic models to cell behavior in the absence of detailed signal-pathway information [24].

Physiological models

The previous section dealt with models that reveal relationships between cell signaling and function. Given the current state of knowledge, these approaches face great challenges in translating their predictions to clinically measurable outcomes. Although breakthroughs in the fields of tissue engineering and the use of more physiologically relevant cell systems have aided in addressing this issue, the problem of understanding physiology through the use of cell-based computational models remains a difficult one. For predictive models of physiology to be feasible, a system has to have been extensively studied or be describable at a high level of abstraction. Given the small number of systems that presently fit these criteria, efforts in this field have produced considerably fewer results than the approaches mentioned earlier in this review. There are, however, a handful of encouraging results that have been published and an increasing number of laboratories and companies are becoming involved in such efforts.

Noble [25] describes a remarkable computational model of the heart that provides a unified description of organ-level physiology in terms of protein-level biology. The model provides nonintuitive explanations for how anti-arrhythmia drugs might work. Extensive knowledge of signaling pathways, cell–cell organization, and the tissue geometry of the heart made this project possible. Related efforts, such as the Physiome Project, attempt to use computation to describe many more systems from the protein level through to the organ level [26]. Other groups have tried to model pathophysiology starting with an organ level (or sometimes even more abstracted) model and then adding layers of information as necessary [27]. For example, mathematical models of type 2 diabetes have been used for over a decade to understand key parameters, such as insulin sensitivity and β-cell function, pertaining to pathology in patients [28]. Entelos has used a top-down modeling approach to generate ODE-based models at the organ level. Lewis *et al.* [29] used this framework to explore the pathophysiology of asthma and found that the resulting model could capture the acute and chronic characteristics of asthma. Importantly, their model also correctly predicted a lack of clinical inefficacy for humanized anti-interleukin-5 antibodies [27]. Although the number of models that allow for this type of powerful prediction are few, it is clear from studies, such as Lewis *et al.*, that the potential benefits for a pharmaceutical company are extremely high. These studies can be used not only to predict the clinical outcome of a particular drug, but also to identify novel interventions that can 'front load' the R&D pipeline with physiologically relevant targets.

Conclusions and future directions in industry

Computational models address a key issue in the pharmaceutical industry: prediction. Whereas many traditional biological studies present insightful descriptions, the application of the hypotheses to new parameter space, such as different ligand concentrations or receptor numbers, is often difficult. Computational modeling solves this problem by bringing existing knowledge together in a quantitative framework that allows scientists to predict the effects of system perturbations. Of course, models do not always predict correctly. When models fail, the underlying assumptions are flawed and fixing this depends on identifying areas that require further experimentation. Companies best able to execute tight integration between modeling and experiment together with repeated iteration of the modeling–experiment cycle will reap the greatest benefit from computational approaches.

Computational models fit within the workflow of the pharmaceutical R&D pipeline, serving to coordinate and explain information being generated in biological and informatics groups. How well these models serve their function will depend on the effective training of scientists that possess both biological intuition and computational skills. The *in silico* component in research must still be coupled with hypothesis-driven experimental design and is not a substitute for the more important *in cerebro* component. The interrogation of models by those who have biological understanding will be vital for the development of successful models. Most large pharmaceutical firms currently have computational modeling groups, including Johnson and Johnson, Eli Lilly, AstraZeneca, Pfizer, Novartis, GlaxoSmithKline and Merck. In addition, several small model-focused companies, such as Gene Network Sciences, Entelos, BG Medicine, BioSeek and Merrimack Pharmaceutical, are developing state-of-the-art computational models for the pharmaceutical industry [30]. We believe that the most successful models will not only provide predictive power but will also be scalable, meaning that models currently appropriate for different phases in the R&D pipeline should be mutually compatible in anticipation of information that will connect disparate R&D stages. Specifically, because PK models have already proven essential for the drug development process, new models capable of integrating with PK models will be most useful for pharmaceutical companies. As new computational tools become available, companies successful in applying modeling will have a competitive advantage derived from increases in predictive power along the R&D pipeline.

Acknowledgements

NK is funded by the NIH Biotechnology Training Grant at MIT. KAJ is a postdoctoral fellow of the American Cancer Society.

References

1 Butcher, E.C. (2005) Can cell systems biology rescue drug discovery? *Nat. Rev. Drug Discov.* 4, 461–467
2 DiMasi, J.A. *et al.* (2003) The price of innovation: new estimates of drug development costs. *J. Health Econ.* 22, 151–185
3 Searls, D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* 4, 45–58
4 Blume-Jensen, P. and Hunter, T. (2001) Oncogenic kinase signalling. *Nature* 411, 355–365
5 Ideker, T. and Lauffenburger, D. (2003) Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol.* 21, 255–262
6 Kholodenko, B.N. (2006) Cell-signalling dynamics in time and space. *Nat. Rev. Mol. Cell Biol.* 7, 165–176
7 Lee, E. *et al.* (2003) The roles of APC and Axin derived from experimental and theoretical analysis of the Wnt pathway. *PLoS Biol.* 1, E10
8 Behrens, J. (2005) The role of the Wnt signalling pathway in colorectal tumorigenesis. *Biochem. Soc. Trans.* 33, 672–675
9 Hendriks, B.S. *et al.* (2006) Computational modeling of ErbB family phosphorylation dynamics in reponse to transforming growth factor alpha and heregulin indicates spatial compartmentation of phosphatase activity. *IEE Proc. Syst. Biol.* 153, 22–33
10 Kim, H.J. and Bar-Sagi, D. (2004) Modulation of signalling by Sprouty: a developing story. *Nat. Rev. Mol. Cell Biol.* 5, 441–450
11 Hatakeyama, M. *et al.* (2003) A computational model on the modulation of mitogen-activated protein kinase (MAPK) and Akt pathways in heregulin-induced ErbB signalling. *Biochem. J.* 373, 451–463
12 Schoeberl, B. *et al.* (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.* 20, 370–375
13 Hendriks, B.S., *et al.* Decreased internalization of ERBB1 mutants in lung cancer is linked with a mechanism conferring sensitivity to gefitinib. *IEE Proc. Syst. Biol.* (in press)
14 Marshall, C.J. (1995) Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. *Cell* 80, 179–185
15 Segarra, J. *et al.* (2006) Combined Signaling through ERK, PI3K/AKT, and RAC1/p38 Is Required for Met-triggered Cortical Neuron Migration. *J. Biol. Chem.* 281, 4771–4778
16 Zheng, P.Z. *et al.* (2005) Systems analysis of transcriptome and proteome in retinoic acid/arsenic trioxide-induced cell differentiation/apoptosis of promyelocytic leukemia. *Proc. Natl. Acad. Sci. U. S. A.* 102, 7653–7658
17 Hirai, M.Y. *et al.* (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* 101, 10205–10210
18 Kao, K.C. *et al.* (2004) Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis. *Proc. Natl. Acad. Sci. U. S. A.* 101, 641–646
19 Hwang, D. *et al.* (2005) A data integration methodology for systems biology. *Proc. Natl. Acad. Sci. U. S. A.* 102, 17296–17301
20 Hwang, D. *et al.* (2005) A data integration methodology for systems biology: experimental verification. *Proc. Natl. Acad. Sci. U. S. A.* 102, 17302–17307
21 Janes, K.A. *et al.* (2005) A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science* 310, 1646–1653
22 Clayton, T.A. *et al.* (2006) Pharmaco-metabonomic phenotyping and personalized drug treatment. *Nature* 440, 1073–1077
23 Christopher, R. *et al.* (2004) Data-driven computer simulation of human cancer cell. *Ann. N. Y. Acad. Sci.* 1020, 132–153
24 Aksenov, S.V. *et al.* (2005) An integrated approach for inference and mechanistic modeling for advancing drug development. *FEBS Lett.* 579, 1878–1883
25 Noble, D. (2006) Systems biology and the heart. *Biosystems* 83, 75–80
26 Hunter, P.J. and Borg, T.K. (2003) Integration from proteins to organs: the Physiome Project. *Nat. Rev. Mol. Cell Biol.* 4, 237–243
27 Butcher, E.C. *et al.* (2004) Systems biology in drug discovery. *Nat. Biotechnol.* 22, 1253–1259
28 Kansal, A.R. (2004) Modeling approaches to type 2 diabetes. *Diabetes Technol. Ther.* 6, 39–47
29 Lewis, A. *et al.* (2001) The roles of cells and mediators in a computer model of chronic asthma. *Int. Arch. Allergy Immunol.* 124, 282–286
30 Mack, G.S. (2004) Can complexity be commercialized? *Nat. Biotechnol.* 22, 1223–1229